Finite-horizon dynamic programming

Tianxiao Zheng
SAIF

1. Introduction

Markov decision processes can be solved by linear programming or dynamic programming. In this note, we present the latter approach. The method of dynamic programming is best understood by studying finite-horizon problems. Therefore, we start with the case of finite horizon and introduce the infinite horizon dynamic programming next. The idea of this method is to use backward induction.

2. Principle of optimality

> An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision. (Bellman)

The principle of optimality states that

1. Dynamic programming principle
   Because the agent does not make decision in the last period, we define value function at time T,

$$V_T(s_T) = u_T(s_T).$$

For $t = 0, 1, ..., T-2, T-1$, the value function can be defined as

$$V_t(s_t) = \max_{\{a_j\}_{j=t}^{T-1}} \quad \mathbb{E}_t \sum_{j=0}^{T-t} \beta^j u_{t+j}(s_{t+j}, a_{t+j}).$$

It can be shown that the value function has the following recursive structure

$$V_t(s_t) = \max_{a_t} \quad u_t(s_t, a_t) + \beta \mathbb{E}_t V_{t+1}(s_{t+1}).$$

The above equation is called *Bellman equation or dynamic programming equation.* For $t = 0$, we reach the original Markov decision problem

$$V_0(s_0) = \max_{\{a_j\}_{j=0}^{T-1}} \quad \mathbb{E}_0 \sum_{t=0}^{T} \beta^t u_t(s_t, a_t).$$

2. Verification theorem
   The solution to the Bellman equation is the value function and that associated policy is the optimal policies.
3. Any optimal policy obtained from solving the Markov decision problem can be generated by solving the Bellman equation.

With the principle of optimality, we are guaranteed that the optimal policy (plan) obtained from the Markov decision problem is identical to those generated from the optimal policy functions obtained from dynamic programming.

3. Solving optimal stopping problem

Here, we give a specific example to show how to solve the optimal stopping problem with the method of dynamic programming.

Example 3.1. *The secretary problem*

State space: $\mathbb{Z} = \{0, 1\}$. 1 denotes that the current one is the best seen so far, and 0 denotes that a previous one was better. Let $V_t(1)$ ($V_t(0)$) be the maximum probability of choosing the best candidate, when the current candidate is (is not) the best among the first t candidates.

At time $t = T = N$, it is obvious that $V_N(0) = h(0) = 0$, $V_N(1) = h(1) = 1$.

For $t = 0, 1, ..., T-2, T-1$, the Bellman equation can be written as

$$\begin{aligned} V_t(z_t) &= max(f_t(z_t), g_t(z_t)) + max\, \mathbb{E}_t V_{t+1}(s_{t+1}) \\ &= max(f_t(z_t), g_t(z_t)) + max\, (\frac{1}{t+1}V_{t+1}(1) + \frac{t}{t+1}V_{t+1}(0), 0) \\ &= max\left(f_t(z_t) + \frac{1}{t+1}V_{t+1}(1) + \frac{t}{t+1}V_{t+1}(0),\ g_t(z_t)\right). \end{aligned} \tag{1}$$

As a result,

$$V_t(0) = \frac{1}{t+1}V_{t+1}(1) + \frac{t}{t+1}V_{t+1}(0) \quad \textit{(always choose to continue)}, \tag{2}$$

$$V_t(1) = max\left(\frac{1}{t+1}V_{t+1}(1) + \frac{t}{t+1}V_{t+1}(0),\ t/N\right) = max(V_t(0),\ t/N). \tag{3}$$

We can see from Eq. 3 that $V_t(1) \geq V_t(0)$, and from Eq. 2 that $V_t(0) \geq V_{t+1}(0)$. Therefore, $V_t(0)$ is decreasing with t. From Eq. 3, we see that there is a critical t_s so that $V_{t_s}(0) = t_s/N$. This means that the optimal policy is to interview the first t_s candidates and to stop at the first one that is the best so far.

For $t > t_s$, we have $V_t(1) = t/N$, and $V_t(0) = \frac{1}{N} + \frac{t}{t+1}V_{t+1}(0)$. With $V_N(0) = 0$, we solve the $V_t(0)$ equation backward and get

$$V_t(0) = \frac{t}{N}(\frac{1}{t} + \frac{1}{t+1} + ... + \frac{1}{N-1}), \quad t \geq t_s$$

This can be used to find t_s,

$$\frac{t_s}{N} = \frac{t_s}{N}(\frac{1}{t_s} + \frac{1}{t_s+1} + ... + \frac{1}{N-1}).$$

For sufficient large N, we could approximate the above equation as,

$$1 \approx \log \frac{N}{t_s}.$$

This is due to the following inequality

$$\int_{t_s}^{N} \frac{1}{\tau} d\tau < \frac{1}{t_s} + \frac{1}{t_s+1} + \cdots + \frac{1}{N-1} < \int_{t_s-1}^{N-1} \frac{1}{\tau} d\tau \tag{4}$$

Therefore $t_s/N \approx 1/e$.

4. Application to optimal control

Modern macroeconomic models often present themselves as a optimal control problem. In these problems, the state s is decomposed into a vector of endogenous state x and a vector of exogenous state z. The exogenous state evolves according to a time-homogeneous Markov process with transition function Q. The endogenous state evolves according to the following equation

$$x_{t+1} = \phi_t(x_t, a_t, z_t, z_{t+1}) \qquad t = 0, 1, ..., T-1,\ (x_0, z_0) \text{ given.}$$

This equation is called state transition equation. The action a_t is a vector of control variables. After choosing $a_t \in \Gamma(x_t, z_t)$, the decision maker obtains reward $u_t(x_t, z_t, a_t)$.

We consider the control problem

$$\max_{\{a_t\}_{t=0}^{T-1}} \mathbb{E}\left[\sum_{t=0}^{T-1} \beta^t u_t(x_t, z_t, a_t) + \beta^T u_T(x_T, z_T)\right].$$

The associated Bellman equation is given by

$$V_t(x_t, z_t) = \max_{a_t \in \Gamma(x_t, z_t)} u_t(x_t, z_t, a_t) + \mathbb{E}_t \beta V_{t+1}(x_{t+1}, z_{t+1}),$$

and $V_T(x_T, z_T) = u_T(x_T, z_T)$.

4.1. Characterization of the value function: monotonicity, concavity and differentiability

The properties of the solution to the above Bellman equation (*i.e.* V_t and policy G_t: $\mathbb{X}^{n_x} \times \mathbb{R}^{n_z} \to \mathbb{R}^{n_a}$) are discussed Stokey, Lucas, and Prescott, 1989. It can be shown that

- under the condition <u>(1)</u> $u_t(., z, a)$ $(t = 0, 1, ...T-1)$ is strictly increasing for each z, a and

$u_T(.,z)$ is strictly increasing for each z, (2) for each z, $\Gamma(.,z)$ is increasing ($x < x'$ implies $\Gamma(x,z) \subset \Gamma(y,z)$), (3) $\phi_t(.,a,z,z')$ is increasing for each a,z,z', *then* $V_t(.,z)$ *is strictly increasing for each* z.

- under the condition (4) $u_T(.,z)$ is strictly concave for each z and at each z, for all x, a, x', a' and $\theta \in (0,1)$, $t = 0,1,...,T-1$, $u_t(\theta x+(1-\theta)x', z, \theta a+(1-\theta)a') \geq \theta u_t(x,z,a)+(1-\theta)u_t(x',z,a')$, (5) at each z, for all x, x', $a \in \Gamma(x,z)$ and $a' \in \Gamma(x',z)$ imply that $\theta a + (1-\theta)a' \in \Gamma(\theta x + (1-\theta)x', z)$, (6) $\phi(.,.,z,z')$ is concave for each z, z', *then* $V_t(.,z)$ *is strictly concave for each* z*;* G_t *is a single-valued continuous function.*
- under (4)-(6), (7) for each z, $u_T(.,z)$ is differentiable on the interior of $\mathbb{X}$ and $u_t(.,z,.)$ ($t = 0,1,...T-1$) is continuously differentiable on the interior of $\mathbb{X}\times\mathbb{A}$, (8) for each z, z', $\phi(.,.,z,z')$ is differentiable on the interior of $\mathbb{X} \times \mathbb{A}$, *then* $V_t(.,z_t)$ *is continuously differentiable.*

Differentiating the Bellman equation with respect to x_t gives the **envelope condition**,

$$\frac{\partial V_t(x_t,z_t)}{\partial x_t} = \frac{\partial u_t}{\partial x_t} + \mathbb{E}_t \beta \frac{\partial V_{t+1}}{\partial x_{t+1}} \frac{\partial \phi_t}{\partial x_t}, \quad \text{for } t = 1,2,...,T-1,$$

and $\frac{\partial V_T(x_T,z_T)}{\partial x_T} = \frac{\partial u_T(x_T,z_T)}{\partial x_T}$. The solution of the Bellman equation can be computed by iterating backward from T to 0 and choosing a_t at $t = T-1, T-2, ..., 1, 0$. To find a_t, we need the **first order condition** of the Bellman equation

$$0 = \frac{\partial u_t}{\partial a_t} + \beta \mathbb{E}_t \frac{\partial V_{t+1}}{\partial x_{t+1}} \frac{\partial \phi_t}{\partial a_t}, \quad \text{for } t = 0,1,...T-1.$$

Below we give an example to illustrate the use of dynamic programming method to solve the optimal control problem.

Example 4.1. *A consumption-saving problem*

Consider a classical consumption-saving problem with uncertain labor income. A consumer is initially endowed with some savings. Each period he receives uncertain labor income. He then decides how much to consume and how much to save in order to maximize his discounted expected utility over a finite horizon.

We formulate this problem by the following Markov decision model:

$$max\, \mathbb{E}\left[\sum_{t=0}^{T} \beta^t u(c_t)\right],$$

subject to

$$c_t + a_{t+1} = Ra_t + y_t, \quad a_0 \geq 0 \text{ given}$$

where $R > 0$ *is the gross interest rate and* c_t*,* a_t*, and* y_t *represent consumption, savings, and labor income, respectively. Assume each* y_t *is independently and identically drawn from a distribution over* $[y_{min}, y_{max}]$*.*

We need to impose a constraint on a_{T+1}*; otherwise, the consumer could borrow and consume an* *infinite amount. Thus, we may impose a borrowing constraint of the form* $a_{T+1} \geq -b_T$ *where* $b_T \geq 0$ *represents a borrowing limit. In general, there may be borrowing constraints in each period because of financial market frictions:* $a_{t+1} \geq -b_t$ *, for some* $b_t \geq 0$, $t = 0, 1, ..., T$, *where the borrowing limit* b_t *may be time and state dependent. Here we simply do not allow borrowing:* $b_t = 0$.

For convenience, we redefine the state variable as cash hold at each date $x_t \equiv Ra_t + y_t$. *The state transition equation becomes* $x_{t+1} = R(x_t - c_t) + y_{t+1}$. *Defining* $\mathcal{A}_t \equiv x_t - c_t = a_{t+1}$ *as the control variable, we transform the original problem to*

$$max\ \mathbb{E}\left[\sum_{t=0}^{T} \beta^t u(x_t - \mathcal{A}_t)\right],$$

subject to $x_{t+1} = R\mathcal{A}_t + y_{t+1}$.

Now we solve the problem by the method of dynamic programming. As in Schechtman and Escudero, 1977, we assume $u(c_t) = -\exp(-\gamma c_t)$, $(\gamma > 0)$.

For $t = T$, $\mathcal{A}_T = 0$, *the value function* $V_T(x_T) = u(x_T)$. *The Bellman equation is*

$$V_t(x_t) = \max_{0 \leq \mathcal{A}_t \leq x_t} \quad u(x_t - \mathcal{A}_t) + \beta \mathbb{E}_t V_{t+1}(R\mathcal{A}_t + y_{t+1}),$$

where the expectation is taken with respect to y. The envelope condition can be written as (x_t does not appear in the state transition equation, resulting in a very simple envelope condition. *This is the reason why we define* x_t *and* $\mathcal{A}_t$ *in the first place.)*

$$V_t'(x_t) = \frac{\partial u(x_t - \mathcal{A}_t)}{\partial x_t}.$$

The first order condition of Bellman equation is

$$0 = \frac{\partial u(x_t - \mathcal{A}_t)}{\partial \mathcal{A}_t} + \beta R \mathbb{E}_t V_{t+1}'(R\mathcal{A}_t + y_{t+1})$$

Substituting the envelope condition into the first order condition, we get

$$e^{-\gamma(x_t - \mathcal{A}_t)} = \beta R \mathbb{E}_t e^{-\gamma(R\mathcal{A}_t + y_{t+1} - \mathcal{A}_{t+1})}, \quad \{t = 0, ..., T-1\}, \text{ and } \mathcal{A}_T = 0$$

This equation is usually called ***Euler equation****.*

We start solving the Euler equation from $t = T - 1$ *(*$\mathcal{A}_T = 0$*) and get*

$$\mathcal{A}_{T-1} = \frac{x_{T-1}}{1+R} + \frac{\log(\beta R)}{\gamma(1+R)} + \frac{\log(\mathbb{E}_{T-1} e^{-\gamma y_T})}{\gamma(1+R)}$$

For $t = T-2$,

$$e^{\gamma(1+R)\mathcal{A}_{T-2}} = \beta R e^{\gamma x_{T-2}} \mathbb{E}_{T-2} \exp\left[\gamma\left(\frac{R\mathcal{A}_{T-2} - Ry_{T-1}}{1+R} + \frac{\log(\beta R)}{\gamma(1+R)} + \frac{\log(\mathbb{E}_{T-1}e^{-\gamma y_T})}{\gamma(1+R)}\right)\right]$$

$$\frac{1+R+R^2}{1+R}\mathcal{A}_{T-2} = x_{T-2} + \frac{\log(\beta R)}{\gamma(1+R)} + \frac{\log(\beta R)}{\gamma} + \frac{\log(\mathbb{E}_{T-1}e^{-\gamma y_T})}{\gamma(1+R)} + \frac{\log \mathbb{E}_{T-2}e^{-\gamma y_{T-1}R/(1+R)}}{\gamma}$$

Observing the structure of $\mathcal{A}_{T-1}$ and $\mathcal{A}_{T-2}$, we conjecture that

$$\mathcal{A}_t = (1-D_t)x_t + C_t\frac{\log(R)}{\gamma} + B_t, \quad \textit{with } D_T = 1, C_T = 0, B_T = 0$$

Substituting the preceding conjecture into the Euler equation yields

$$-\gamma x_t + \gamma\mathcal{A}_t = \log(\beta R) - \gamma R\mathcal{A}_t + \gamma(1-D_{t+1})R\mathcal{A}_t + C_{t+1}\log(\beta R) + \gamma B_{t+1} + \log \mathbb{E}_t e^{-\gamma D_{t+1}y_{t+1}}$$

i.e.

$$(1+D_{t+1}R)\left((1-D_t)x_t + C_t\frac{\log(R)}{\gamma} + B_t\right) = x_t + (C_{t+1}+1)\frac{\log(\beta R)}{\gamma} + B_{t+1} + \frac{1}{\gamma}\log \mathbb{E}_t e^{-\gamma D_{t+1}y_{t+1}}$$

Matching coefficients yields

$$\begin{aligned} D_{t+1}R(1-D_t) &= D_t \\ C_t(1+D_{t+1}R) &= C_{t+1} + 1 \\ B_t(1+D_{t+1}R) &= B_{t+1} + \frac{1}{\gamma}\log \mathbb{E}_t e^{-\gamma D_{t+1}y_{t+1}} \end{aligned} \tag{5}$$

Then we get

$$\begin{aligned} D_t &= \frac{RD_{t+1}}{1+RD_{t+1}} \\ C_t &= (1-D_t)(1+C_{t+1}) \\ B_t &= (1-D_t)B_{t+1} + (1-D_t)\frac{1}{\gamma}\log \mathbb{E}_t e^{-\gamma D_{t+1}y_{t+1}} \end{aligned} \tag{6}$$

Iterating the relations (with $D_T = 1, C_T = 0, B_T = 0$) gives D_t, C_t, B_t, e.g.

$$D_{T-1} = \frac{R}{1+R}, \quad C_{T-1} = \frac{1}{1+R}, \quad B_{T-1} = \frac{1}{1+R}\frac{1}{\gamma}\log \mathbb{E}_{T-1}e^{-\gamma y_T},$$

$$\begin{aligned} D_{T-2} &= \frac{R^2/(1+R)}{1+R^2/(1+R)} = \frac{R^2}{1+R+R^2}, \quad C_{T-2} = \frac{1+R}{1+R+R^2}\left(1+\frac{1}{1+R}\right) \\ B_{T-2} &= \frac{1+R}{1+R+R^2}\frac{\log(\mathbb{E}_{T-1}e^{-\gamma y_T})}{\gamma(1+R)} + \frac{1+R}{1+R+R^2}\frac{1}{\gamma}\log \mathbb{E}_{T-2}e^{-\gamma\frac{R}{1+R}y_{T-1}}, \quad \textit{and so on ...} \end{aligned} \tag{7}$$

and the optimal saving policy $\mathcal{A}_t = (1-D_t)x_t + C_t\frac{\log(\beta R)}{\gamma} + B_t$. The optimal consumption policy is *given by $c_t = x_t - \mathcal{A}_t$.*

4.2. Maximum principle

The optimal control problem can also be solved by the maximum principle.

Maximum Principle is originally proposed for continuous-time optimal control problems by Pontryagin *et al.* (1962). This principle is closely related to the Lagrange multiplier method and also applies to discrete time optimal control problems. It is pretty interesting to study the maximum principle and its relation to dynamic programming.

We start by write the Lagrangian form for the control problem

$$\mathcal{L} = E\left[\sum_{t=0}^{T-1} \beta^t u_t(x_t, z_t, a_t) - \beta^{t+1}\mu_{t+1}(x_{t+1} - \phi_t(x_t, z_t, a_t, z_{t+1})) + \beta^T u_T(x_T, z_T)\right]$$

We can derive the first-order conditions for the solution. Specifically, differentiating with respect to a_t yields

$$\frac{\partial u_t}{\partial a_t} + \beta \mathbb{E}_t \mu_{t+1} \frac{\partial \phi_t}{\partial a_t} = 0$$

for $t = 0, 1, ..., T-2, T-1$. If μ_t is interpreted as the shadow value of the value function $\mu_t = \frac{\partial V_t(x_t, z_t)}{\partial x_t}$, we immediately recognize that the above equation is the first order condition of the Bellman equation.

Differentiating with respect to x_t yields,

$$\mu_t = \frac{\partial u_t}{\partial x_t} + \beta \mathbb{E}_t \mu_{t+1} \frac{\partial \phi_t}{\partial x_t}$$

for $t = 1, 2, ..., T-1$ and $\mu_T = \frac{\partial u_T}{\partial x_T}$ for $t = T$. If μ_t is interpreted as the shadow value of the value function $\mu_t = \frac{\partial V_t(x_t, z_t)}{\partial x_t}$, we immediately recognize that the above equation is the envelope condition of the Bellman equation.

Combining the first-order conditions with the state-transition equation yields $n_x \times T + n_x \times T + n_a \times T$ system equations

$$\begin{aligned} 0 &= \frac{\partial u_t}{\partial a_t} + \beta \mathbb{E}_t \mu_{t+1} \frac{\partial \phi_t}{\partial a_t}, \\ \mu_t &= \frac{\partial u_t}{\partial x_t} + \beta \mathbb{E}_t \mu_{t+1} \frac{\partial \phi_t}{\partial x_t}, \quad (\mu_T = \frac{\partial u_T}{\partial x_T}), \\ x_{t+1} &= \phi_t(x_t, a_t, z_t, z_{t+1}), \quad x_0 \text{ given}, \end{aligned} \tag{8}$$

for the same number of variables $(a_t, x_{t+1}, \mu_{t+1})_{t=0}^{T-1}$. These system equations give the necessary conditions for optimality and can be solved recursively by backward induction. Suppose we can use the first of the system equations to solve for a_t as a function of x_t, z_t, and μ_{t+1}. Substituting a_t into the other two system equations, we obtain a system of two first order difference equation for $\{x_{t+1}, \mu_{t+1}\}_{t=0}^{T-1}$. We need two boundary conditions to obtain the solution. Luckily, we have them:

μ_T and x_0. (In the infinite horizon case, μ_T is replaced by a transversality condition.)

Example 4.2. *The consumption-saving problem solved by maximum principle*

The Lagrangian form for the control problem is given by

$$\mathcal{L} = \mathbb{E}\left[\sum_{t=0}^{T-1} \beta^t u(x_t - \mathcal{A}_t) - \beta^{t+1}\mu_{t+1}(x_{t+1} - R\mathcal{A}_t - y_{t+1}) + \beta^T u(x_T)\right]$$

Differentiating with respect to $\mathcal{A}_t$ yields

$$\frac{\partial u(x_t - \mathcal{A}_t)}{\partial \mathcal{A}_t} + \beta R \mathbb{E}_t \mu_{t+1} = 0$$

for $t = 0, 1, ..., T-2, T-1$. Differentiating with respect to x_t yields,

$$\mu_t = \frac{\partial u(x_t - \mathcal{A}_t)}{\partial x_t}$$

for $t = 1, 2, ..., T-1$ and $\mu_T = \frac{\partial u}{\partial x_T}$ for $t = T$.

The system equations are then given by

$$\begin{aligned} \mu_t &= \gamma e^{-\gamma(x_t - \mathcal{A}_t)},\ \mu_T = \gamma e^{-\gamma x_T},\ \{t = 1, ..., T-1\} \\ 0 &= -\gamma e^{-\gamma(x_t - \mathcal{A}_t)} + \beta R \mathbb{E}_t \mu_{t+1},\ \{t = 0, ..., T-1\} \\ x_{t+1} &= R\mathcal{A}_t + y_{t+1},\ \text{given } x_0,\ \{t = 0, ..., T-1\} \end{aligned}$$

By simple algebra, we get from the system equations

$$e^{-\gamma(x_t - \mathcal{A}_t)} = \beta R \mathbb{E}_t e^{-\gamma(R\mathcal{A}_t + y_{t+1} - \mathcal{A}_{t+1})}, \quad \{t = 0, ..., T-1\},\ \text{and } \mathcal{A}_T = 0.$$

This is just the Euler equation obtained using the method of dynamic programming.

The advantage of Maximum principle is that we do not need to study value functions directly which are complicated objects as shown in the previous section. Instead, the Lagrange multipliers are the objects to solve for. In addition, this approach can easily handle additional *intra-temporal* constraints on states or actions. One only needs to introduce additional Lagrange multipliers and then apply the Kuhn-Tucker Theorem to derive first-order conditions.

Appendix A. Another form of Lagrangian

In many economic problems, the state-transition equation takes the form $x_{t+1} = \phi(x_t, a_t, z_t)$. In this case, we can replace the Lagrange multiplier $\beta^{t+1}\mu_{t+1}$ with $\beta^t\lambda_t$. The first order conditions become

$$\frac{\partial u_t}{\partial a_t} + \lambda_t \frac{\partial \phi_t}{\partial a_t} = 0$$

for $t = 0, 1, ..., T-2, T-1$ and

$$\lambda_t = \beta \mathbb{E}_t \frac{\partial u_{t+1}}{\partial x_{t+1}} + \beta \mathbb{E}_t \lambda_{t+1} \frac{\partial \phi_{t+1}}{\partial x_{t+1}}$$

for $t = 0, 1, ..., T-2$ and $\lambda_{T-1} = \beta \mathbb{E}_{T-1} \partial u_T(x_T, z_T)/\partial x_T$.

References

Schechtman, J., Escudero, V. L., 1977. Some results on an income fluctuation problem. Journal of Economic Theory 16, 151 – 166.

Stokey, N. L., Lucas, R. E., Prescott, E. C., 1989. Recursive Methods in Economic Dynamics. Harvard University Press.